Canada

Catalogue no. 81-595-MIE — No. 005

Education, skills and learning – Research papers

Linking provincial student assessments with national and international assessments

by Fernando Cartwright, Diane Lalancette, Jerry Mussio, and Dehui Xing

Culture, Tourism and the Centre for Education Statistics Division
2001 Main Building, Ottawa, K1A 0T6
Telephone: 1 800 307-3382 Fax: 1 613 951-9040

This paper represents the views of the authors and does not necessarily reflect the opinions of Statistics Canada, or the British Columbia Ministry of Education

Statistics Canada Statistique Canada

British Columbia Ministry of Education
Ministère de l'Éducation de la Colombie-Britannique

105-03911/

Education, skills and learning

Research papers

Linking provincial student assessments with national and international assessments

Fernando Cartwright, *Statistics Canada*
Diane Lalancette, *British Columbia Ministry of Education*
Jerry Mussio, *Statistics Canada*
Dehui Xing, *British Columbia Ministry of Education*

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2003

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without prior written permission from Licence Services, Marketing Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

May 2003

Catalogue no. 81-595-MIE2003005

Frequency: Irregular

ISSN 1704-8885

ISBN 0-662-34213-5

Ottawa

La version française de cette publication est disponible sur demande (nº 81-595-MIF2003005 au catalogue)

Statistics Canada

British Columbia Ministry of Education

This paper represents the views of the authors and does not necessarily reflect the opinions of Statistics Canada, or the British Columbia Ministry of Education

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Client Services, Culture, Tourism and the Centre for Education Statistics, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: (613) 951-7608; toll free at 1 800 307-3382; by fax at (613) 951-9040; or e-mail: educationstats@statcan.ca).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our Web site.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
E-mail inquiries	infostats@statcan.ca
Web site	www.statcan.ca

Ordering information

This product, Catalogue No. 81-595-MIE2003005, is available on the Internet for free. Users can obtain single issues at: http://www.statcan.ca/cgi-bin/downpub/studiesfree.cgi.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136.

Acknowledgements

We would like to thank the students, teachers and principals who participated in the 2000 OECD PISA study and the 2000 British Columbia Foundation Skills Assessment (FSA). We also gratefully acknowledge staff from the Student Assessment and Program Evaluation Branch of the British Columbia Ministry of Education and from the Centre for Education Statistics at Statistics Canada for their on-going support of this project. The contributions of Valerie Collins and Markus Baer of the British Columbia Ministry, who assisted in the analysis of instruments, are especially appreciated. Sincere thanks are also extended to Danielle Baum of Statistics Canada for her help in preparing the manuscript for publication.

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Table of Contents

Summary		5
1.	Introduction	6
2.	Overview of FSA and PISA	8
3.	What FSA and PISA tell us about reading achievement in British Columbia	11
4.	Linking FSA with PISA	15
5.	Results	17
6.	Conclusions	22
Appendix A:	Tables	23
Appendix B:	Comparison of FSA and PISA	27
Appendix C:	Technical methodology	29
References		39
Endnotes		40

Catalogue no. 81-595-MIE2003005

Summary

The purpose of this feasibility study was to develop technical procedures that will enable ministries of education to link provincial tests with national and international tests so that standards can be compared and results reported on a common scale.

The technical procedures were developed and used to link reading tests administered by British Columbia's annual Foundation Skills Assessment (FSA) and the Programme for International Student Assessment (PISA). For this feasibility study, we examined linkages between the FSA reading test administered to the population of grade 10 students in spring 2000 and the PISA reading test administered about the same time to a sample of 15-year-olds.

Results of linking the two assessments show that the FSA standard to recognize reading excellence in British Columbia is set higher than the PISA standard used to identify top readers across 32 countries. If the equivalent of this international standard were used to identify excellent readers by FSA, a greater number of grade 10 B.C. students would be classified as excellent readers in the annual provincial assessment. For example, FSA in 2000 reported that 9% of B.C. grade 10 students exceeded expectations. About double that number would have been classified as exceeding expectations if FSA used the equivalent of the PISA standard to identify excellent readers.

The linkage also makes it possible to look at the performance of 15-year-old students in other educational jurisdictions in terms of the B.C. assessment. For example, Finland and Alberta had the highest PISA scores among all countries and provinces in 2000, and their strong performance is reflected when their international scores are reported in terms of the B.C. reading scale. Specifically, 81% of 15-year-olds from Finland meet or exceed B.C. reading standards, followed by Alberta 15-year-olds at 80%. Proportions of students meeting or exceeding standards are about the same for Quebec and B.C., 77% and 76% respectively. Across Canada, 75% of 15-year-olds meet or exceed B.C. reading standards.

Among G-7 countries, the percentage of 15-year-olds meeting or exceeding the B.C. reading standard ranges from lows of 58% in Germany, 59% in Italy and 64% in the United States, to a high of 75% in Canada and Japan.

Estimates of the proportion of students from another jurisdiction who meet or exceed FSA reading standards are typically accurate, across jurisdictions, within three percentage points 19 times out of 20.

Provinces can use the methodology developed for this study to compare provincial standards with national and international standards and to report results on a common scale. The methodology can be used to establish linkages between two assessments when a common sample of students completes both tests or when random equivalent samples of students are selected. The procedures are valid for linking test scores for groups of students (greater than 30, for most statistics); they are not appropriate for linking and reporting scores for individual students.

1. Introduction

Most ministries of education in Canada carry out province-wide testing in elementary and secondary schools. These assessments enable ministries to report student performance in relation to standards set by their provinces.

Results from provincial, national and international assessments are difficult to compare because test results are reported on different scales.

In addition, all ministries now routinely participate in pan-Canadian and international assessments of student learning. Provincial governments participate in these external assessments to compare performance of their students with those in other provinces and countries. For policy makers across Canada, these assessments have become important tools for judging how well students are prepared to participate in a global knowledge society.

While provincial and international assessments are now routinely administered in Canadian schools, reports produced by provincial governments often appear to conflict with those produced by international agencies. It is not unusual, for example, for the media to report student "failure rates" of 30% based on the latest provincial test, and then a few months later report the findings of an international assessment showing the province scoring among the "best in the world". Both reports may be correct – but can be confusing to those who assume the results are being reported on the same scales or that the two assessments have measured the same phenomenon.

Results from different assessments can be compared if their scales are linked to a common scale.

Even when assessments can be shown to assess the same phenomenon (e.g. reading proficiency), comparing results from provincial tests with those reported by international assessments is not unlike trying to compare daily temperature reports from capital cities around the world without knowing which scale is being used in different countries, Fahrenheit or Celsius.

The purpose of this feasibility study was to establish technical procedures that could enable ministries of education to link provincial tests with national and international tests so that test results may be reported on a common scale. Where instruments are linked, a province can determine if its standards are higher or lower than those set by national or international bodies; it can also report the performance of other jurisdictions in relation to provincial standards.

The ability to compare results from large-scale assessments is becoming increasingly important as educators and the public alike try to reconcile perceived differences between provincial, national and international reports on student achievement.

Catalogue no. 81-595-MIE2003005

There is another reason for linking assessments. Large-scale assessments are expensive to develop and administer, and they consume considerable time and energy in schools. Establishing linkages between provincial and international tests holds the promise of improving the richness and cost-effectiveness of provincial assessment programs by making it possible to incorporate international benchmarks in routine provincial assessment reports.

This study examines a linkage between B.C.'s Foundation Skills Assessment (FSA) and the Programme for International Student Assessment (PISA).

For this feasibility study, we examine a linkage between the reading tests administered by British Columbia's Foundation Skills Assessment (FSA) and by the Programme for International Student Assessment (PISA) carried out by the Organization for Economic Cooperation and Development (OECD). This study uses reading assessment data collected as part of FSA 2000 and PISA 2000, both of which were administered in the period April to May, 2000. At that time in British Columbia, 2800 grade 10 students participated in both FSA and PISA, providing a good opportunity to link the two assessment scales.

Technical procedures used in the study were developed at Statistics Canada and build on the work carried out by other researchers.

In the spring of 2000, 2800 grade 10 students in B.C. participated in both FSA and PISA, providing a good opportunity to link the two assessment scales.

This report summarizes results from the feasibility study. Part 2 of the report provides background information about FSA and PISA. Part 3 displays test results for B.C. students as reported by FSA and PISA. Part 4 provides an overview of the procedures used to link the two assessments. Part 5 provides results of the linkage procedure, and conclusions are presented in Part 6.

2. Overview of FSA and PISA

About FSA

FSA assesses reading, writing and numeracy, and is administered annually to all eligible B.C. students enrolled in grades 4, 7 and 10.

The Foundation Skills Assessment is an annual province-wide assessment providing a snapshot of how well students in British Columbia are attaining skills in reading, writing, and numeracy. The skills tested are linked to the provincial curriculum with a focus on skills required to function in school and everyday life.

FSA is administered every spring to the population of grade 4, 7 and 10 students in public and provincially funded independent schools. Results are issued in the early fall of the following school year. The first complete assessment was administered in 2000 and serves as the benchmark year for comparisons in reading and numeracy. For writing, the benchmark year is 2001.

FSA results are used by the B.C. school system to monitor student achievement and to develop plans for improvement.

The main purpose of FSA is to assist schools, school districts and the ministry in evaluating how well students are achieving basic skills and in developing plans to improve student achievement. A secondary purpose is to provide parents with additional information about their child's progress.

FSA is designed and administered by the Student Assessment and Program Evaluation Branch of the British Columbia Ministry of Education. Teachers and other educators are extensively involved in the development of test items, scoring of open-response test items, and setting of standards.

FSA reading scores are generated using item response theory. Results are then reported in terms of three performance standards: exceeds expectations, meets expectations, not yet within expectations (see Box 1). Test scores required for each of these performance standards were established by the ministry in 2000 based on advice from panels of experts. A sample of common test items is used to link assessments from one year to the next so that FSA results for each year can be reported on a common scale. The Ministry plans to review the FSA standards periodically.

BOX 1

FSA Standards of Performance

Exceeds expectations: The level of a student's performance that is beyond that at which a teacher would say the student has fully met the expectations of the grade on this test. Student performance would be considered excellent for the grade on this test.

Meets expectations: The level of performance at which a student meets the widely held expectations for the grade on this test.

Not yet within expectations: With no other information, this is the level at which a teacher would want to know more about the reasons for a student's low performance.

Catalogue no. 81-595-MIE2003005

Summary results, which are provided to schools and to the public at large, display the proportion of students within each performance level. Other statistical data, such as average scores, are also reported to assist in the interpretation of data, but the focus of attention is on the proportion of students who meet or exceed provincial expectations.

FSA public reports place emphasis on the proportion of students meeting or exceeding provincial standards.

FSA results are combined with other indicators, such as report card data and student retention rates, to provide key information for growth plans developed by school planning councils and for accountability contracts prepared by school districts. Schools provide parents with individual FSA results for their children along with relevant interpretative information.

About PISA

The Programme for International Student Assessment (PISA) is a collaborative effort among member countries of the Organisation for Economic Co-operation and Development (OECD). PISA assesses the performance of 15-year-olds in three domains: reading literacy, mathematical literacy and scientific literacy. The term "literacy" is used to reflect emphasis on skills required to function in society.

PISA assesses reading, mathematics and science; it is being administered in 2000, 2003 and 2006 to samples of 15-year-olds in 32 countries.

Three PISA cycles have been planned – 2000, 2003, and 2006 – each one focussing on a different literacy domain. In 2000, the major focus was reading literacy, with mathematical and scientific literacy as minor domains. As a result, there were fewer mathematics and science items included in the assessment and these items were administered to a sub-sample of participants. Mathematical and scientific literacy will be the focus in 2003 and 2006, respectively.

PISA assesses how far students near the end of compulsory education have acquired some of the knowledge and skills that are essential for full participation in society. It presents evidence on student performance in reading, mathematical and scientific literacy, and reveals factors that influence the development of these skills at home and at school. These results are intended to be used to inform policy development.

The purpose of PISA is to inform policy development across OECD countries.

As with FSA, PISA reading scores are generated using item response theory (although the two assessments use different variations of the item response model). PISA reading results are reported using two main statistics: average scores and the proportion of students attaining each of five reading proficiency levels (see Box 2). Standard cut points for each proficiency level are defined by subject experts drawn from participating countries and measurement specialists representing the OECD. As with B.C.'s assessment program, changes in PISA scores are measured over time, with 2000 used as the benchmark year.

Thirty-two countries participated in PISA 2000. In Canada, approximately 30,000 15-year-old students from more than 1,000 schools took part. A large Canadian sample was drawn so that results could be reliably reported at both national and provincial levels; there are no reports for individual schools or students.

The PISA 2000 survey also included questionnaires to collect background information from students and school principals. In Canada, students' parents were also contacted by telephone survey as part of the Youth in Transition Survey. Students

participating in this survey will be contacted every two years; a key purpose of this study is to identify factors that influence student success in further education and in the job market.

BOX 2

PISA Reading Proficiency Levels

Students proficient at **Level 5** (over 625 points) are capable of completing sophisticated reading tasks, such as managing information that is difficult to find in unfamiliar texts; showing detailed understanding of such texts and inferring which information in the text is relevant to the task; and being able to evaluate critically and build hypotheses, draw on specialised knowledge, and accommodate concepts that may be contrary to expectations.

Students proficient at **Level 4** (553 to 625 points) are capable of difficult reading tasks, such as locating embedded information, construing meaning from nuances of language and critically evaluating a text.

Students proficient at **Level 3** (481 to 552 points) are capable of reading tasks of moderate complexity, such as locating multiple pieces of information, drawing links between different parts of the text, and relating it to familiar everyday knowledge.

Students proficient at **Level 2** (408 to 480 points) are capable of basic reading tasks, such as locating straightforward information, making low-level inferences of various types, deciding what a well-defined part of the text means, and using some outside knowledge to understand it.

Students proficient at **Level 1** (335 to 407 points) are capable of completing only the least complex reading tasks developed for PISA, such as locating a single piece of information, identifying the main theme of a text or making a simple connection with everyday knowledge.

Students performing **below Level 1** (below 335 points) are not able to show routinely the most basic type of knowledge and skills that PISA seeks to measure. These students have serious difficulties in using reading literacy as an effective tool to advance and extend their knowledge and skills in other areas.

Summary results are released to OECD member countries and to global media. Average scores for each country are displayed along with the proportion of students attaining each reading proficiency level. In Canada, PISA 2000 results for each province and key demographic groups were shared at the same time as international results were announced.

PISA results are used in various ways by participating countries. In Germany, PISA 2000 results sparked a comprehensive review of the German education system. In Canada, provincial ministries for the most part combine PISA results, and other pan-Canadian and international test data, with provincial test data when reporting educational progress to the public.

Catalogue no. 81-595-MIE2003005

3. What FSA and PISA tell us about Reading achievement in B.C.

FSA Results

FSA 2000 reports that 79% of B.C. grade 10 students meet or exceed provincial reading standards and 21% of students do not meet provincial standards.

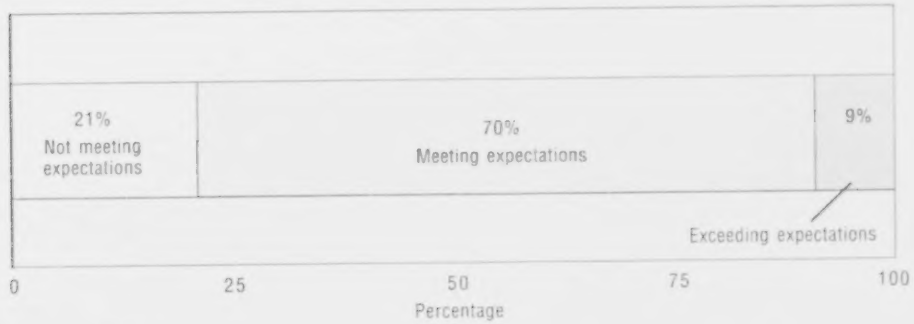
In the spring of 2000, students in grades 4, 7 and 10 completed the FSA tests. All students were expected to participate; guidelines for the exemption of some students with special needs were provided to school principals who were responsible for decisions regarding exemptions. Almost all students (94%) in grades 4 and 7 completed FSA reading tests, while 82 % of grade 10 students completed the reading test. The FSA population is defined by the students who were not exempt from the assessment, and the students assessed are considered to be a census.

The focus of this linkage study is the FSA reading test administered to grade 10 students in the spring of 2000. FSA reported that 79% of grade 10 students met or exceeded provincial expectations and 21% of students did not meet provincial expectations.

Figure 1

Percentage of B.C. grade 10 students attaining provincial reading standards

Reported by the Foundation Skills Assessment (FSA), 2000



Source: Table 1.

FSA 2000 summary results were issued to the 60 school districts in the province, and to all public and funded independent schools. In addition to reporting results for the population of students, the B.C. Ministry published test results for various student subgroups and highlighted variations in performance. The reading performance of boys and of aboriginal students has been of particular concern in British Columbia. FSA 2000 reported that 74% of grade 10 males meet or exceed the reading standard compared to 85% of grade 10 females. Fifty-eight percent of

aboriginal students meet or exceed the reading standard compared to 79% for the population as a whole.

Between 2000 and 2002, the proportion of grade 10 students meeting or exceeding FSA reading standards declined from 79% to 71%. The proportion of grade 7 students meeting or exceeding reading expectations also declined, from 81% to 76% over the three-year period. At grade 4, this proportion improved slightly from 79% to 80% In writing and numeracy, the proportion of students meeting or exceeding expectations increased in relation to the benchmark years for all three grades (Table 1).

PISA Results

PISA 2000 reports that the average reading score for B.C. 15-year-olds is at about the same level as those from top-scoring Alberta and Finland.

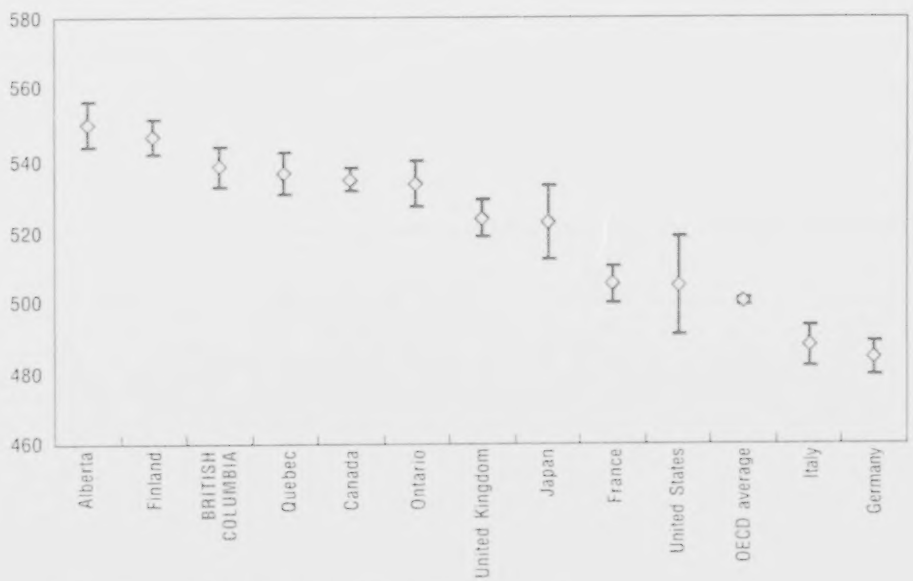
Average scores: When average test scores for OECD countries are compared, Canada's 15-year-olds rank very high in reading literacy compared to their counterparts in the 32 countries participating in PISA 2000. Only Finland had higher average scores than Canada. New Zealand, Australia, Ireland and Japan performed at about the same level as Canada. (Table 2).

British Columbia students also scored high relative to other jurisdictions in the world. Figure 2 displays average PISA Reading scores for B.C. 15 year-olds compared to their counterparts in selected provinces and in G-7 countries[1]. Finland has been included in the comparison as it scored highest among OECD countries in the reading component of PISA 2000.

Figure 2

Average PISA Reading scores of 15-yr-olds in British Columbia and selected jurisidictions

Reported by the Programme for International Student Assessment (PISA) 2000



Average (mean) scores are bounded by 95% confidence intervals.

Source: Table 2.

When sampling errors are taken into account, B.C. average scores are about the same as those from top scoring Alberta and Finland, and about the same as Quebec, Ontario and Japan. Average scores for British Columbia students are higher than those from the United Kingdom, France, United States, Italy, Germany, and the OECD average.

As with B.C.'s reading assessment, PISA reported that average reading scores for 15-year-old females were higher than their male counterparts—this same finding was reported for every country participating in PISA. Another key finding compares average scores of students from families with the highest socio-economic status (SES) with those of students from families with the lowest SES. Compared to other countries, Canada and Finland are similar in that they combine high overall PISA reading scores with relatively small achievement differences by SES.

Proficiency levels: Ranking average scores tells us little about what students can actually do. For this reason, the reading achievement scale in PISA 2000 was divided into five proficiency levels (Box 2).

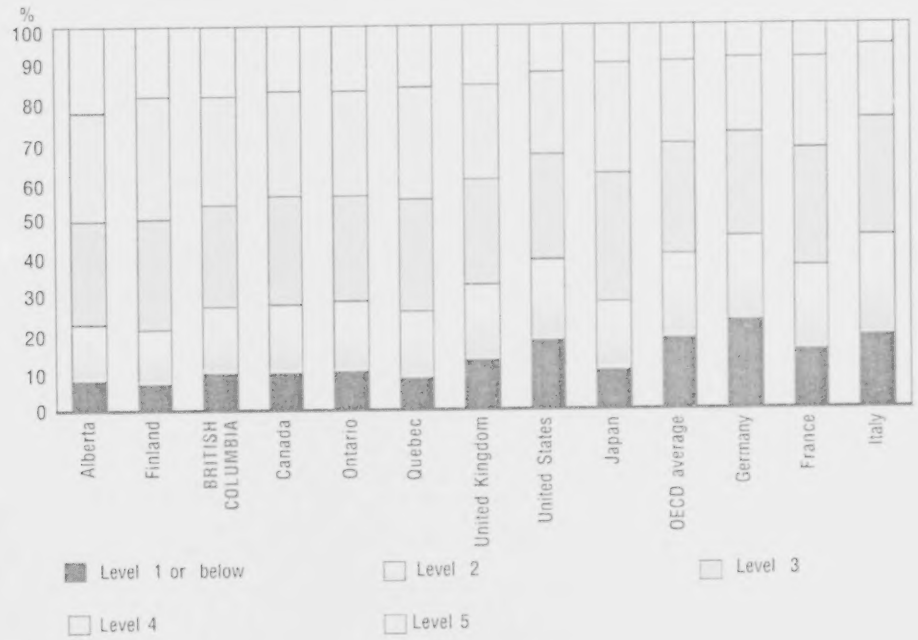
Figure 3 displays reading skill profiles for selected provinces and countries. Alberta, at 23%, and Finland, at 19%, placed a high proportion of students at the top reading level, with just 8% and 7% at the lowest reading levels, respectively. British Columbia, at 18%, also placed a high proportion of students at the top reading level with 9% scoring at the lowest levels.

PISA 2000 also reports that 18% of B.C. students place at the highest international reading level and 9% at the lowest level.

Figure 3

Percentage of 15-yr-olds attaining PISA Reading Proficiency Levels, B.C. and selected jurisdictions

Reported by the Programme for International Student Assessment (PISA) 2000



Jurisdictions ordered by percentage of students attaining Level 5.

Source: Table 2.

To the extent that strong reading skills can be considered predictive of a better-skilled citizenry and work force, countries and jurisdictions with a high proportion of students attaining high reading levels will have an important social and economic advantage. On the other hand, those young people with poor, or low-level, literacy skills may find it difficult to benefit from further educational opportunities and may be limited in their ability to contribute and participate in societies dependent on information and knowledge.

Summary

FSA 2000 reports that 79% of grade 10 students meet or exceed B.C. reading standards and 21% do not. PISA 2000 reports that B.C. 15-year-olds – who are about six months younger on average than grade 10 students—score high compared to their counterparts in other countries. PISA also reports that 18% of B.C. 15-year-olds score at the top international reading standard and 9% at the lowest levels, findings that, again, are very good internationally. So are B.C. reading standards higher or lower than the international benchmarks set by PISA? Can these provincial and international data be reported on a common scale so that the results can be better understood?

Catalogue no. 81-595-MIE2003005

4. Linking FSA with PISA

Comparing instruments

FSA 2000 (grade 10 reading) and PISA 2000 (reading) were compared to determine if the same reading domain is being measured.

Establishing linkages between FSA and PISA reading tests is possible only if the same skill or content area is being measured. If two tests measure different domains of reading or if they measure the same domains differently, then students are likely to exhibit different proficiency patterns on the tests; in these cases scores on one test will not provide accurate estimates of scores on the second test.

Students taking part in FSA were assessed on their capacity to identify and interpret key concepts and main ideas; locate, interpret, and organize details; and analyze what they read critically. Students taking part in PISA were assessed on their capacity to retrieve specified information; interpret what they read; and, reflect and evaluate what they read, drawing on their existing knowledge.

FSA used a single booklet of 41 reading items administered to all students; separate test booklets for writing and numeracy were administered during the same week. PISA used 129 items to assess reading. These items were divided into nine booklets and each student in the sample responded to one booklet; eight of the booklets also included items in mathematics and science.

Three reading assessment specialists with extensive experience in the design of FSA were asked to review each of the 41 FSA test items, classify each item according to the PISA assessment framework, and identify items that could not be classified. They then undertook the reverse procedure: they were asked to review the 129 PISA test items and link each item to the FSA assessment framework using the same criteria. (See Appendix A)

The reviewers concluded that the two assessments measure similar reading skills.

The reviewers reported that all of the 41 FSA test items were classified within the PISA assessment framework. In the reverse process, all of the 129 PISA items were also classified within the FSA framework; the reviewers observed, however, that 11 of the PISA items reflected a greater emphasis on the application of reading skills than was evident in the FSA criteria. Overall, the specialists reported that both test instruments measure similar reading skills. Having established that the two assessments measure approximately the same phenomena, statistical procedures were then used to link the scales on which the results were reported.

Statistical model

Establishing the mathematical relationship linking FSA and PISA reading scales is analogous to determining the relationship between Fahrenheit and Celsius temperature scales.

The next step in the linkage process involves the development of mathematical models linking the two assessments. The temperature analogy helps explain what is involved in linking two test score scales. By analyzing repeated measures of air temperature using both Fahrenheit and Celsius thermometers, we can establish the mathematical relationship $C = (5/9)(F-32)$. This formula tells us that a temperature

of 86 on the Fahrenheit scale is equivalent to 30 on the Celsius scale. In a similar way, we can analyze measures of student achievement using two instruments – in this case PISA and FSA—and use the data to develop the mathematical relationship between the two scales.

Measuring a cognitive activity such as reading, however, is more difficult than measuring temperature, and the formula linking FSA with PISA is more complex.

The temperature example is only a partial analogy because measuring temperature is very simple compared to the measurement of complex cognitive activities such as reading or mathematics. While errors can occur if poor thermometers are used, temperature is measured directly and the formula linking Celsius and Fahrenheit scales is exact. Measuring a cognitive activity such as reading is subject to error because what is being measured is often defined in different ways and the measurement procedure, usually a paper and pencil test, is an indirect one.

Linking FSA scores with PISA scores requires that three sources of error be quantified: measurement error, linking error, and sampling error. Measurement error describes the uncertainty of students' original scores and is a property of the original test instruments. Linking error describes the uncertainty in transforming a score on an original scale into a score equivalent on another scale. This error is expressed on the new scale, and is a property of the mathematical procedure used to estimate the score equivalents – the linking function. Sampling error describes the uncertainty of statistics used to estimate characteristics (e.g. averages and proportions) of the population from which a sample was drawn. Much of the technical work in this project focussed on quantifying the first two errors and incorporating these when reporting the results of the linking function. The methods used to combine all three sources of error for reporting sample statistics is described in Appendix C.

Three techniques can be used to link two assessments: a sample of common test items is included in both instruments; a sample of students completes both tests; or randomly equivalent samples of students are selected. The methods developed for this study may be applied to either of the latter two linking techniques.

The linkage carried out in this study was based on test results for the sample of 2800 grade 10 students who completed both the FSA test and the PISA test over a one-month period in the spring of 2000. Even though individual students may change in reading proficiency during this period, the methods used here assume that the distribution of proficiency is similar; using this assumption, statistical models were developed to link the two assessment scales. (See Appendix C for a summary of the technical approach used to develop the linkage formulae).

In summary, errors in interpretation will result if there are significant differences in what was measured, how the assessments were administered, and the reliability of the two instruments. To the extent the two tests measure the same skills, were administered under the same conditions to the same population and have small errors of measurement, the error associated with linking two tests will be small and the inferences drawn will have strong validity.

Catalogue no. 81-595-MIE2003005

5. Results

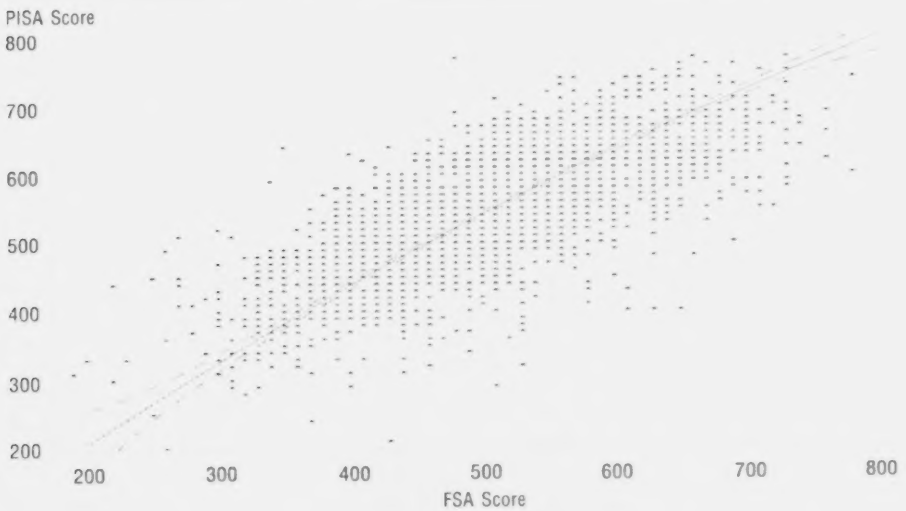
Mathematical functions were developed linking the two assessments; one was developed linking PISA to FSA, and a second was developed for the reverse linkage, FSA to PISA. The two functions are inverse. For example, applying one function to a FSA score will produce the PISA equivalent, and applying the other function to this equivalent will recover the original FSA score.

The mathematical relationship between FSA and PISA is represented in Figure 4.

The mathematical equation for expressing the linkage is not as simple as that for the conversion of Celsius to Fahrenheit, because of the complexity of the procedures used to estimate the relationship. The non-linear relationship between the FSA scales and the PISA equivalents is illustrated in Figure 4.

Figure 4

Relationship between FSA and PISA test scores[1]



1. Scores have been rounded to the nearest 10. Each point may represent several cases.
The dashed lines represent the 68% confidence interval of the FSA to PISA linking function.

The linking function is displayed against a scatter plot representing the score pairs of the students in the linking sample, although the estimation method did not use regression methods or the individual score pairs (see Appendix C). The dashed lines represent the standard error of the FSA to PISA linking function. Table 3 in Appendix A shows a sample of linked PISA scores and FSA scores, along with their associated linking errors. For low and larger scores on either scale, the larger errors indicate greater uncertainty about the value of the score equivalences. For the

majority of the observed score ranges of both scales, the score equivalents are expected to be accurate within nine points 19 times out of 20.

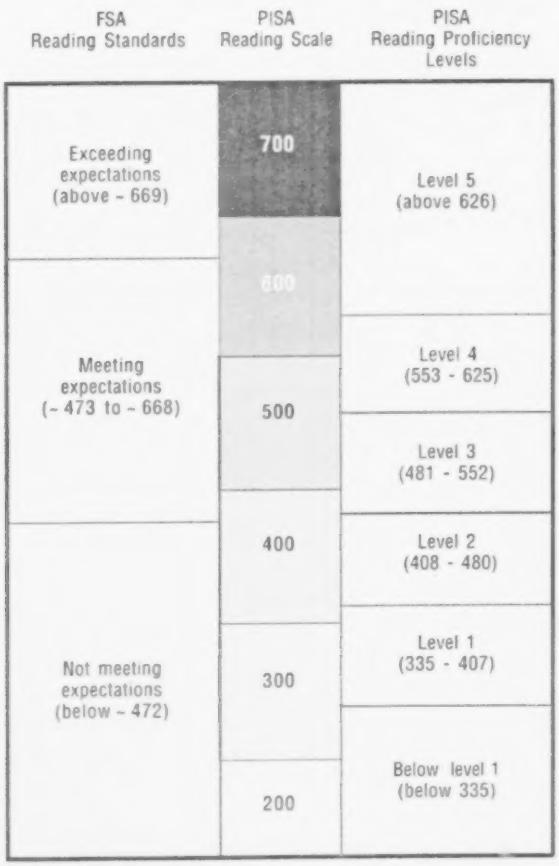
The small dissimilarities between the two assessments (noted above) suggest that the linked scores of one test are not interchangeable with scores from the other test for indivudal students[2]. However, the assumptions underlying the linking procedures require that averages of linked scores for groups of students be interchangeable with averages of original scores. Therefore, linking FSA and PISA scores is valid when the results are used to make inferences about groups of students (about 30 or more) but not for linking and reporting scores for individuals.

Comparing reading standards

Once the two scales were linked, the scores defining FSA's reading benchmarks were converted to PISA scores, and vice versa (Tables 4, 5). FSA and PISA benchmarks are compared in Figure 5 using the PISA international scale.

Figure 5

Comparing FSA and PISA standards using the PISA reading scale



Source: Tables 4, 5.

The FSA benchmark to identify reading excellence is set well above the threshold for Level 5, the highest PISA reading proficiency level.

Two observations can be made when FSA and PISA standards are compared. First, the threshold for the highest FSA performance level "Exceeds expectations", which recognizes excellence in reading, is set well above the threshold for PISA Level 5, the highest PISA reading proficiency level. This means that if the equivalent of PISA Level 5 were used as the standard to identify top readers by FSA, a greater number of B.C. students would be classified each year as exceeding provincial expectations. For example, FSA 2000 reported that 9% of B.C. grade 10 students exceeded expectations. If FSA had used the equivalent of PISA Level 5 to identify excellent readers, approximately double that number would have been classified as exceeding expectations in 2000.

If FSA used the equivalent of PISA Level 5 to identify excellent readers, a greater number of B.C. grade 10 students would be classified as Exceeding expectations in the annual provincial assessment.

Second, reading scores categorised as "Meets expectations" in FSA cover a wide range of reading difficulty – Levels 3, 4, 5 as defined by PISA. This suggests that B.C. students who are classified as "Meets expectations" are capable of reading tasks of moderate complexity (Level 3), but may also be capable of more sophisticated reading tasks (Levels 4, 5).

B.C. students who are classified as "Not within expectations" are performing at about PISA Level 2 or below. This suggests that these students are capable of basic reading tasks and making low-level inferences (Level 2), they are capable of completing only the least complex reading tasks (Level 1), or they have serious difficulties in using reading literacy as an effective tool in learning (Below Level 1).

Detailed descriptions of the skills associated with different PISA reading proficiency levels are provided by OECD (2001).

BOX 3

FSA Minimum score: Meets expectations	<— about the same as —>	PISA Minimum score: Level 3
The level of performance at which a student meets the widely held expectations for the grade on this test.		Students are capable of reading tasks of moderate complexity, such as locating multiple pieces of information, drawing links between different parts of the text, and relating it to familiar everyday knowledge.

As noted earlier, reading standards established for both FSA and PISA reflect the professional judgments of panels of experts. British Columbia or any other province may choose, for a variety of reasons, to set standards that are higher or lower than those set by other provinces or by national or international agencies such as the OECD.

At what level standards are set will depend on the purposes of testing. The reality is that achievement standards for most provincial assessments are usually set without knowing if these standards are higher or lower than those set for similar tests by other jurisdictions. If provincial assessments are periodically linked with external tests, a province may compare standards and make refinements with a more complete knowledge base.

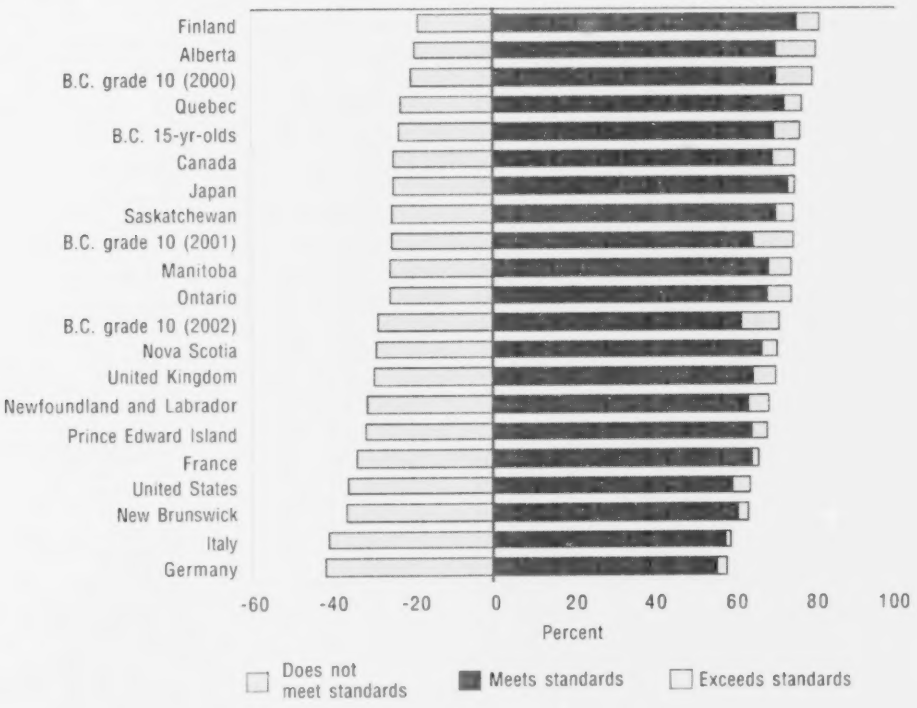
Performance by jurisdictions in relation to B.C.'s reading standard

Linking provincial assessments with external tests enables a province to imbed national and international benchmarks in its provincial assessment reports.

Linking with external tests enables a province to report performance levels for other provinces and countries in relation to its provincial standards. Figure 6 displays the findings when PISA scores are linked to B.C.'s FSA. Results for each jurisdiction are expressed in relation to B.C.'s reading standard.

Figure 6

Percentage of 15-yr olds from various jurisdictions attaining B.C. grade 10 reading standards, 2000



1. All results shown here are for 15-year-olds except for B.C. grade 10 students who are, on average, 6 months older than B.C. 15 year olds.

Jurisdictions ordered by the percentage of students meeting or exceeding expectations.

Source: Table 6

81% of 15-year-olds from Finland and 80% from Alberta meet or exceed B.C. reading standards compared to 77% of students from Quebec and 76% from B.C..

Finland and Alberta had the highest PISA scores among all countries and provinces in 2000, and their strong performance also shows when their international scores are reported in terms of the B.C. reading scale. Specifically, 81% of 15-year-olds from Finland meet or exceed B.C. reading standards, followed by Alberta 15-year-olds at 80%. Proportions of students meeting or exceeding standards are about the same for Quebec and B.C., 77% and 76% respectively. Across Canada, 75% of 15-year-olds meet or exceed B.C. reading standards.

Catalogue no. 81-595-MIE2003005

Among G-7 countries, the percentage of 15-year olds meeting or exceeding the B.C. reading standard range from lows of 58% in Germany, 59% in Italy and 64% in the United States, to a high of 75% in Canada and Japan.

64% of 15-year-olds in the U.S. meet or exceed B.C. reading standards compared to 75% in Canada and in Japan.

Table 6 includes the standard errors associated with statistics produced when PISA scores are transformed to their equivalent FSA scores. These standard errors combine measurement error, linking error and the PISA sampling error. Estimates of the proportion of students from another jurisdiction who meet or exceed FSA reading standards are typically accurate, across provinces and countries, within three percentage points 19 times out of 20.

Some caution is required when interpreting these results. This linkage was estimated using a sample drawn from the population of 15-year-olds in Grade 10 in British Columbia; as the systematic differences between this linking population and the regions to which these results are generalised increase, the validity of interpretations decreases. The countries in which PISA was administered are diverse across many social, economic, and cultural indicators. Strong inferences may be made for populations that are similar to the population for which the linkage was estimated, such as students in other Canadian provinces or similarly developed countries. However, results of the linkage for populations that are very dissimilar from the linking sample should be interpreted as rough approximations rather than precise estimates.

For B.C. students enrolled in grade 10 in 2000, FSA reported that 79% meet or exceed the B.C. reading standard; this proportion declined to 75% in 2001 and to 71% in 2002. B.C. grade 10 students are on average about six months older than B.C. 15-year-olds.

It was noted above that the FSA standard recognizing reading excellence is set higher than the international PISA standard used to identify top readers. As a result, FSA reports fewer students at the top reading level (exceeding expectations). For example, 9% of B.C. 15-year olds score in the top reading category on FSA, while 18% score in the top PISA reading level. For Alberta students, 10% score in the top B.C. category and 23% in the top PISA reading level.

Although not reported here, the linkage formulae can also be applied so that FSA scores for B.C. schools and school districts can be displayed in terms of PISA reading proficiency levels.

6. Conclusions

Provincial, pan-Canadian and international assessments are now routinely administered in Canadian schools. For policy makers across Canada, these assessments have become important tools for judging how well students are prepared to participate in a global knowledge society.

It has been difficult, however, to compare results from the various assessment programs because different reporting scales are used. It is not unusual for the media to report high failure rates on a provincial test, and then several months later report favourable and seemingly contradictory results from an international assessment. The purpose of this feasibility study was to develop technical procedures that will enable ministries of education to link provincial tests with pan-Canadian and international tests so that standards of different assessments can be compared and results reported on a common scale.

Reporting test results on a common scale will make it easier for the public to understand how well students in the province are performing in relation to provincial and international benchmarks. Assessments are expensive to design and administer, so linking these instruments and improving the richness of assessment reports will also help improve the cost-effectiveness of provincial assessment programs.

The technical procedures were developed and used to link reading tests administered by B.C.'s Foundation Skills Assessment (FSA) and the Programme for International Student Assessment (PISA). FSA reading standards were compared to PISA standards. PISA results for other provinces and countries were converted to the B.C. scale and reported in relation to B.C.'s reading standards.

The methodology developed here can be used to establish linkages between two assessments when a common sample of students completes both tests or when randomly equivalent samples of students are selected. Other linking methods can be used when assessments are designed to share common items or are administered simultaneously. The computer programs developed for this study can be used to link scales for assessments that use a variety of scaling methods, including item response theory and classical test theory.

Linkages of this type are valid for linking test scales when the results are used to make inferences about groups of students (about 30 or more); they are not appropriate for linking and reporting scores for individuals.

Catalogue no. 81-595-MIE2003005

Appendix A: Tables

Table 1

Performance of British Columbia Students on FSA, Reading, Writing and Numeracy[1,2,3]
Foundation Skills Assessment, 2000-2002

		Not yet meeting expectations	Meets expectations	Exceeds expectations	Meets or exceeds expectations	Standard error
Reading						
Grade 4	2000	20.5	71.6	7.9	79.5	(0.10)
	2001	22.1	72.8	5.1	77.9	(0.09)
	2002	20.1	73.5	6.4	79.9	(0.10)
Grade 7	2000	19.3	72.9	7.9	80.7	(0.10)
	2001	24.2	66.4	9.4	75.8	(0.10)
	2002	23.6	64.5	11.9	76.4	(0.10)
Grade 10	2000	20.6	70.5	8.9	79.4	(0.09)
	2001	25.3	64.7	10.0	74.7	(0.09)
	2002	28.8	61.7	9.6	71.2	(0.12)
Writing						
Grade 4	2001	9.3	89.9	0.9	90.7	(0.13)
	2002	6.2	93.6	0.2	93.8	(0.11)
Grade 7	2001	19.0	77.9	3.1	81.0	(0.14)
	2002	15.6	82.8	1.6	84.4	(0.13)
Grade 10	2001	13.8	81.9	4.3	86.2	(0.13)
	2002	13.2	79.5	7.3	86.8	(0.14)
Numeracy						
Grade 4	2000	20.5	71.1	8.4	79.5	(0.09)
	2001	16.1	73.4	10.5	83.9	(0.09)
	2002	14.6	71.9	13.6	85.4	(0.10)
Grade 7	2000	20.3	70.2	9.5	79.7	(0.09)
	2001	18.8	73.8	7.4	81.2	(0.10)
	2002	17.8	72.8	9.4	82.2	(0.11)
Grade 10	2000	25.3	65.9	8.7	74.7	(0.10)
	2001	23.3	67.1	9.5	76.7	(0.09)
	2002	23.9	64.9	11.2	76.1	(0.13)

1. Standard errors reported here reflect measurement errors of the FSA instruments.
2. No sampling error is included as it is assumed that the population of students participated in FSA.
3. For reading and numeracy, FSA uses an underlying achievement scale that has a mean of 500 and standard deviation of 100.

Table 2

PISA Reading literacy: Means and proficiency levels
Mean scores and percentage of students at each level of proficiency on the combined reading literacy scale
(Standard errors are shown in parentheses)

Jurisdictions[1,2]	Mean		Below Level 1		Level 1		Level 2		Level 3		Level 4		Level 5	
Alberta	550	(3.3)	1.8	(0.5)	6.1	(0.7)	14.7	(0.8)	26.7	(1.2)	28.2	(1.0)	22.5	(1.4)
Finland	546	(2.6)	1.7	(0.5)	5.2	(0.4)	14.3	(0.7)	28.7	(0.8)	31.6	(0.9)	18.5	(0.9)
British Columbia	538	(2.9)	2.4	(0.5)	7.0	(0.7)	17.5	(0.9)	26.3	(1.1)	28.7	(1.0)	18.1	(1.1)
Quebec	536	(3.0)	2.0	(0.4)	6.4	(0.6)	17.2	(0.9)	29.4	(1.1)	29.2	(1.1)	15.9	(1.0)
CANADA	534	(1.6)	2.4	(0.3)	7.2	(0.3)	18.0	(0.4)	28.0	(0.5)	27.7	(0.6)	16.8	(0.5)
Ontario	533	(3.3)	2.6	(0.6)	7.4	(0.5)	18.2	(0.8)	27.5	(0.9)	27.6	(1.1)	16.7	(1.0)
Manitoba	529	(3.5)	2.0	(0.4)	8.6	(0.9)	18.7	(1.2)	29.6	(1.5)	25.2	(1.2)	15.9	(1.2)
Saskatchewan	529	(2.7)	2.0	(0.5)	7.3	(0.5)	19.2	(0.9)	29.8	(1.3)	27.8	(1.1)	14.0	(1.0)
New Zealand	529	(2.8)	4.8	(0.5)	8.9	(0.5)	17.2	(0.9)	24.6	(1.1)	25.8	(1.1)	18.7	(1.0)
Australia	528	(3.5)	3.3	(0.5)	9.1	(0.8)	19.0	(1.1)	25.7	(1.1)	25.3	(0.9)	17.6	(1.2)
Ireland	527	(3.2)	3.1	(0.5)	7.9	(0.8)	17.9	(0.9)	29.7	(1.1)	27.1	(1.1)	14.2	(0.8)
Korea	525	(2.4)	0.9	(0.2)	4.8	(0.6)	18.6	(0.9)	38.8	(1.1)	31.1	(1.2)	5.7	(0.6)
United Kingdom	523	(2.6)	3.6	(0.4)	9.2	(0.5)	19.6	(0.7)	27.5	(0.9)	24.4	(0.9)	15.6	(1.0)
Japan	522	(5.2)	2.7	(0.6)	7.3	(1.1)	18.0	(1.3)	33.3	(1.3)	28.8	(1.7)	9.9	(1.1)
Nova Scotia	521	(2.3)	2.9	(0.4)	9.2	(0.9)	20.7	(1.2)	29.0	(1.3)	24.6	(1.5)	13.6	(0.9)
Newfoundland and Labrador	517	(2.8)	3.5	(0.5)	10.3	(0.9)	21	(1.3)	28.4	(1.4)	23.5	(1.2)	13.3	(0.9)
Prince Edward Island	517	(2.4)	2.4	(0.5)	10.4	(1.2)	21.9	(1.2)	28.3	(1.5)	23.9	(1.6)	13.1	(1.1)
Sweden	516	(2.2)	3.3	(0.4)	9.3	(0.6)	20.3	(0.7)	30.4	(1.0)	25.6	(1.0)	11.2	(0.7)
Belgium	507	(3.6)	7.7	(1.0)	11.3	(0.7)	16.8	(0.7)	25.8	(0.9)	26.3	(0.9)	12.0	(0.7)
Austria	507	(2.4)	4.4	(0.4)	10.2	(0.6)	21.7	(0.9)	29.9	(1.2)	24.9	(1.0)	8.8	(0.8)
Iceland	507	(1.5)	4.0	(0.3)	10.5	(0.6)	22.0	(0.8)	30.8	(0.9)	23.6	(1.1)	9.1	(0.7)
Norway	505	(2.8)	6.3	(0.6)	11.2	(0.8)	19.5	(0.8)	28.1	(0.8)	23.7	(0.9)	11.2	(0.7)
France	505	(2.7)	4.2	(0.6)	11.0	(0.8)	22.0	(0.8)	30.6	(1.0)	23.7	(0.9)	8.5	(0.6)
United States	504	(7.1)	6.4	(1.2)	11.5	(1.2)	21.0	(1.2)	27.4	(1.3)	21.5	(1.4)	12.2	(1.4)
New Brunswick	501	(1.8)	5.1	(0.5)	11.7	(0.8)	23.1	(1.2)	29.7	(1.1)	21.0	(1.0)	9.5	(0.6)
OECD average	500	(0.6)	6.2	(0.4)	12.1	(0.4)	21.8	(0.4)	28.6	(0.4)	21.8	(0.4)	9.4	(0.4)
Denmark	497	(2.4)	5.9	(0.6)	12.0	(0.7)	22.5	(0.9)	29.5	(1.0)	22.0	(0.9)	8.1	(0.5)
Switzerland	494	(4.3)	7.0	(0.7)	13.3	(0.9)	21.4	(1.0)	28.0	(1.0)	21.0	(1.0)	9.2	(1.0)
Spain	493	(2.7)	4.1	(0.5)	12.2	(0.9)	25.7	(0.7)	32.8	(1.0)	21.1	(0.9)	4.2	(0.5)
Czech Republic	492	(2.4)	6.1	(0.6)	11.4	(0.7)	24.8	(1.2)	30.9	(1.1)	19.8	(0.8)	7.0	(0.6)
Italy	487	(2.9)	5.4	(0.9)	13.5	(0.9)	25.6	(1.0)	30.6	(1.0)	19.5	(1.1)	5.3	(0.5)
Germany	484	(2.5)	9.9	(0.7)	12.7	(0.6)	22.3	(0.8)	26.8	(1.0)	19.4	(1.0)	8.8	(0.5)
Liechtenstein	483	(4.1)	7.6	(1.5)	14.5	(2.1)	23.2	(2.9)	30.1	(3.4)	19.5	(2.2)	5.1	(1.6)
Hungary	480	(4.0)	6.9	(0.7)	15.8	(1.2)	25.0	(1.1)	28.8	(1.3)	18.5	(1.1)	5.1	(0.8)
Poland	479	(4.5)	8.7	(1.0)	14.6	(1.0)	24.1	(1.4)	28.2	(1.3)	18.6	(1.3)	5.9	(1.0)
Greece	474	(5.0)	8.7	(1.2)	15.7	(1.4)	25.9	(1.4)	28.1	(1.7)	16.7	(1.4)	5.0	(0.7)
Portugal	470	(4.5)	9.6	(1.0)	16.7	(1.2)	25.3	(1.0)	27.5	(1.2)	16.8	(1.1)	4.2	(0.5)
Russian Federation	462	(4.2)	9.0	(1.0)	18.5	(1.1)	29.2	(0.8)	26.9	(1.1)	13.3	(1.0)	3.2	(0.5)
Latvia	458	(5.3)	12.7	(1.3)	17.9	(1.3)	26.3	(1.1)	25.2	(1.3)	13.8	(1.1)	4.1	(0.6)
Luxembourg	441	(1.6)	14.2	(0.7)	20.9	(0.8)	27.5	(1.3)	24.6	(1.1)	11.2	(0.5)	1.7	(0.3)
Mexico	422	(3.3)	16.1	(1.2)	28.1	(1.4)	30.3	(1.1)	18.8	(1.2)	6.0	(0.7)	0.9	(0.2)
Brazil	396	(3.1)	23.3	(1.4)	32.5	(1.2)	27.7	(1.3)	12.9	(1.1)	3.1	(0.5)	0.6	(0.2)

1. Jurisdictions are ordered by mean scores.

2. The PISA reading scale has an international mean of 500 and a standard deviation of 100.

Sources: PISA Canada (2001).
OECD (2001).

Table 3

Linking FSA scores and PISA scores

Linking FSA to PISA			Linking PISA to FSA		
FSA Score	PISA Score	SE	PISA Score	FSA Score	SE
200	194.75	(40.33)	200	206.53	(38.45)
250	235.44	(24.00)	250	267.56	(15.66)
300	277.30	(14.87)	300	326.48	(8.06)
350	320.48	(9.67)	350	383.28	(5.09)
400	365.08	(6.66)	400	437.97	(3.82)
450	411.26	(4.91)	450	490.56	(3.28)
500	459.19	(3.91)	500	541.05	(3.18)
550	509.08	(3.40)	550	589.47	(3.41)
600	561.17	(3.27)	600	635.82	(4.05)
650	615.75	(3.59)	650	680.14	(5.31)
700	673.19	(4.79)	700	722.42	(7.65)
750	733.96	(8.66)	750	762.70	(12.01)
800	798.64	(25.51)	800	801.02	(20.34)

Table 4

FSA benchmarks on PISA scale

FSA Benchmark		PISA Equivalent	
Benchmark	Score	Score	SE
Meets expectations	432.91	472.82	(3.41)
Exceeds expectations	636.88	668.71	(4.90)
Mean	500.00	541.05	(3.18)

Table 5

PISA benchmarks on FSA scale

PISA Benchmark		FSA Equivalent	
Benchmark	Score	Score	SE
Performance Level 1	335	307.38	(10.94)
Performance Level 2	408	372.36	(6.32)
Performance Level 3	481	440.76	(4.23)
Performance Level 4	553	512.14	(3.38)
Performance Level 5	626	589.22	(3.36)
Mean	500	459.19	(3.91)

Table 6

Percentage of 15-year olds students attaining FSA 2000 grade 10 reading standards[1]

	Not yet meeting standards	SE	Meets standards	SE	Exceeds standards	SE	Meets or exceeds standards	SE
Provinces								
Newfoundland and Labrador	31.37	(1.53)	63.63	(1.40)	5.00	(1.17)	68.63	(1.56)
Prince Edward Island	31.87	(1.59)	64.23	(1.51)	3.90	(1.12)	68.13	(1.62)
Nova Scotia	29.07	(1.47)	66.85	(1.43)	4.08	(1.01)	70.93	(1.50)
New Brunswick	36.65	(1.20)	60.76	(1.07)	2.59	(0.83)	63.35	(1.22)
Quebec	23.19	(1.42)	72.45	(1.26)	4.36	(0.77)	76.81	(1.44)
Ontario	25.70	(1.41)	68.13	(1.20)	6.17	(0.93)	74.30	(1.43)
Manitoba	25.46	(1.63)	68.84	(1.44)	5.70	(1.12)	74.54	(1.65)
Saskatchewan	25.22	(1.40)	70.23	(1.30)	4.56	(0.97)	74.79	(1.42)
Alberta	19.70	(1.40)	70.45	(1.32)	9.85	(1.15)	80.30	(1.42)
British Columbia	23.65	(1.32)	70.12	(1.22)	6.23	(0.97)	76.35	(1.34)
British Columbia grade 10 (2000)	20.58	na	70.55	na	8.87	na	79.42	(0.01)
British Columbia grade 10 (2001)	25.25	na	64.73	na	10.02	na	74.75	(0.09)
British Columbia grade 10 (2002)	28.76	na	61.69	na	9.55	na	71.24	(0.12)
Countries								
CANADA	24.73	(0.68)	69.40	(0.58)	5.87	(0.43)	75.27	(0.69)
Finland	18.87	(1.08)	75.82	(0.94)	5.31	(0.71)	81.13	(1.09)
France	33.97	(1.56)	64.21	(1.48)	1.82	(0.63)	66.03	(1.00)
Germany	41.92	(1.26)	55.60	(1.17)	2.48	(0.68)	58.08	(1.27)
Italy	41.06	(1.55)	57.90	(1.46)	1.03	(0.63)	58.94	(1.56)
Japan	24.94	(2.46)	73.42	(2.32)	1.64	(0.66)	75.06	(2.46)
United Kingdom	29.38	(1.07)	64.64	(1.04)	5.98	(0.86)	70.62	(1.08)
United States	35.88	(2.79)	59.76	(2.41)	4.37	(0.92)	64.13	(2.80)

1. All results shown here are for 15-year-olds except for B.C. grade 10 students who are on average 6 months older than B.C. 15 year olds.

Catalogue no. 81-595-MIE2003005

Appendix B: Comparing FSA and PISA test designs

1. Key Features

B.C.'s Foundation Skills Assessment (FSA)[3]

Purpose

The main purpose of the assessment is to help the province, school districts, schools, and school planning councils evaluate how well reading, writing, and numeracy are being addressed and make plans to improve student achievement.

A secondary purpose is to provide teachers, students, and parents with an external source of information about individual student performance.

FSA Reading Comprehension

FSA measures foundation skills that are embedded in the provincial curricula. Although not confined to any single course or grade, the skills assessed by the FSA are most closely linked to prescribed learning outcomes in language arts (and mathematics).

Reading Comprehension is based on the interaction between readers and texts, and assessed through a constructivist, meaning-making approach.

Task Characteristics

Students taking part in FSA were asked questions based on a variety of genres: literary passages, poetry, and informational texts.They were assessed on their capacity to identify and interpret key concept and main ideas, on whether they could locate, interpret, and organize details, and on how well they do critical analysis.

The test included multiple-choice and open-ended tasks.

Reporting Results

An overall score showing reading performance was generated for both multiple-choice and open-ended items using Item Response Theory (two-parameter logistic model). On the basis of these scores, each student was assigned to one of three performance standard: 'Not within expectations', 'Meets expectations', and 'Exceeds expectations'. Individual results are aggregated at the school, district, and provincial level to show percentage of students for each performance levels. Individual student results are reported to students and parents.

OECD Program for International Student Assessment (PISA)[4]

Purpose

PISA assesses how far students near the end of compulsory education (15-year-olds) have acquired some of the knowledge and skills that are essential for full participation in society. It presents evidence on student performance in reading, mathematical and scientific literacy, reveals factors that influence the development of these skills at home and at school, and examines what the implications are for policy development

PISA Reading Literacy

PISA assessed young's people's ability to use their knowledge and skills to meet real-life challenges, rather than looking merely at how well they have mastered a specific school curriculum.

Reading Literacy is defined as understanding, using, and reflecting on written texts, in order to achieve one's goals, to develop one's potential, and to participate in society.

Task Characteristics

Students taking part in PISA were asked questions based on a variety of written texts, ranging from a short story to a letter on the Internet and information presented in a diagram.They were assessed on their capacity to retrieve specified information, on whether they could interpret what they read, and on how well they could reflect and evaluate it, drawing on their existing knowledge.

The test included multiple-choice and open-ended items.

Reporting Results

An overall score showing reading performance was generated for both multiple-choice and open-ended items using Item Response Theory (one-parameter Rasch model). On the basis of these scores, each student was assigned to one of five reading levels. Individual results are aggregated to show the percentage of students who are proficient at each level of reading proficiency and to compare average performances for various populations and demographic groups. Results are reported for 32 countries and ten provinces in Canada. No individual student results are reported.

2. FSA 2000 and PISA 2000 Reading Tasks – A comparison

	Number of items (%) in FSA 2000		Number of items (%) in PISA 2000	
FSA 2000 Reading Tasks				
Identify and interpret key concept and main ideas	4	(9%)	18	(14%)
Locate, interpret, and organize details	24	(56%)	64	(50%)
Critical analysis (make inferences and draw conclusions based on information from the text)	15	(35%)	47	(36%)
Total	43		129	
PISA 2000 Reading Tasks				
Retrieve: locate information in a text	22	(51%)	36	(28%)
Interpret: construct meaning and draw inferences from written information	20	(47%)	63	(49%)
Reflect and Evaluate: relate a text to knowledge, ideas and experiences	1	(2%)	30	(23%)
Total	43		129	

3. Classification of reading text as defined in each assessment

Type of Texts	FSA % of total items
Narrative	42
Poetry	26
Informational	32
Total	100

Type of Texts	PISA % of total items
Advertisements	1
Argumentative/persuasive	13
Charts/Graphs	11
Descriptive	9
Expository	23
Forms	6
Injunctive	7
Maps	3
Narrative	14
Schematics	4
Tables	9
Total	100

Catalogue no. 81-595-MIE2003005

Appendix C

Technical Methodology

The following summary of the procedures examined and developed for linking and estimating statistics for this study is excerpted from Cartwright (in press):

1. Overview

The primary assumption of scale linking methods is that, across repeated measurements, the expected scores of an individual or a test population for one assessment can estimate the expected score for a second assessment. Therefore, the percentile ranks of individuals (i.e., how well they perform, compared to other individuals in the same population) should remain constant across assessments, because percentile rank is a function of the distribution of the underlying latent trait in a given population, rather than the scale of a particular assessment. Furthermore, if two samples are randomly equivalent, the distribution of the latest trait should also be equivalent. Consequently, when randomly equivalent groups are administered two different assessments, the scores corresponding to a particular percentile rank on the different assessments can be used to estimate each other. This method is known as *equipercentile linking*. A complete development of the equipercentile linking function is described in Braun and Holland (1982), where it is described as *equipercentile equating*. Other methods may be more appropriate when common items are used to link assessments.

The steps involved in equipercentile linking are:

a. Estimate cumulative distribution functions (CDFs) for each test score distribution

$$F(x) = Pr(X>x), \qquad (1)$$

where x is a possible test score from the distribution of test scores, X. The CDFs describe the percentile rank corresponding to each observed score.

b. Find inverse CDFs, $F^{-1}(x)$, for each test to predict the observed score given the cumulative distribution (i.e., percentile rank), such that

$$F^{-1}(F^{-1}(x)) = F(x); \qquad (2)$$

c. Use the CDFs and inverse CDFs to find the equipercentile linking functions

$$e_y x = G^{-1}(F(x)), \qquad (3)$$

where $e_y x$ is the equipercentile equivalent of a score on scale Y, given a score x and G is the linking function. Equations 1, 2, and 3, applying to X and x, also apply to Y and y.

The estimation of Equation 1 can be carried out explicitly, but this will produce a "jagged" CDF that is composed of linear segments joining CDF values for each observed score. Because the underlying continuum of proficiency represented by the test scores is believed to be continuous, any function describing the distribution of scores is expected to also be continuous. Therefore, any departures from continuity are assumed to be a result of our inability to sample individuals or measure scores in the missing intervals. However, an empirical CDF that linearly interpolates between observed scores has more discontinuities than there are observed score values. If a better sample had been taken or finer measurement precision were available, presumably the missing intervals would be "filled in" to produce a non-jagged function. Typically, the sample is limited and the instruments are fixed, so the observed CDF must be "smoothed" to produce a continuous CDF, simulating what the distribution of scores would look like with an infinite sample and an assessment of infinite length. The approach used in this study was to estimate continuous score probability density functions (PDFs) and use the integrals of these as continuous CDF estimates. The continuous CDFs are then matched using the steps defined above.

For the current study, several statistical methods of estimating PDFs were compared: (1) Gaussian kernel estimation; (2a) finite mixture models with variable number of components; (2b) finite mixture models with fixed number of components; and (3) 4-parameter Beta distributions. This comparison was necessary due to the uncertain nature of the "true" distribution (i.e., given infinite sample and test length). Each method is appropriate for different types of distributions, but all must be compared in order to determine the most appropriate method for a specific set of data. Based on a statistical comparison of the results for each method and the computing resources required for their implementation, the function used to link the PISA and FSA scales was estimated using 4-parameter Beta distributions. However, these methods represent a spectrum of sampling robustness to data sensitivity. Together, all four methods can be applied as a battery to find the best continuous estimates for a variety of continuous, discrete, non-normal, and multi-modal score distributions. Each of these methods is described in further detail below.

2. Estimating Cumulative Density Functions

In order to prevent undue influence of extreme FSA scores on the estimation of PDFs, cases with scores of 0 and 791.07 on the FSA (the minimum and maximum observed scores, respectively) were removed from the sample used for this project, resulting in a common sample of n = 2659 students. This was done to reduce to influence of measurement error on the estimation procedure. Measurement error for scaled scores produced using item response theory (IRT) typically increases

dramatically at the extremes for tests of finite length (Lord, 1980), reaching a maximum for examinees with uniformly correct or incorrect responses. The most likely proficiencies to produce perfectly correct or incorrect (uniform) response vectors are infinitely high or low, respectively, yet it is unlikely that any examinees truly have infinitely high or low proficiencies. Therefore, point estimation of scores in IRT requires the specification of arbitrary maxima and minima that usually correspond to two or three standard deviations from the mean. Examinees with extreme responses (all correct or all incorrect) receive the same arbitrary high or low scale score. This treatment results in artificially high frequencies for the highest and lowest scores, although it is expected that, had the test more information, these extreme modes would not exist. Score estimation for PISA, which draws random *plausible values* from each examinee's posterior score density instead of single point estimate, eliminated the occurrence of local artificial modes (*PISA 2000 Technical Report,* 2002).

Gaussian Kernel Density Estimation

The Gaussian kernel density estimation method involved in the present study used estimates of the population distributions of scores as the mean of J kernels centred on each of the j observed scores in the sample (Härdle, 1990; Silverman, 1986). Thus, the population density function, $P(x_j)$ is defined by:

$$P(x_i) = \sum_j K(x_i \mid x_j)\, p(x_j), \tag{4}$$

$$K(x_i \mid x_j) = e^{-2^{-1}(x_i - x_j)^2 (h^* s_j)^{-2}}, \tag{5}$$

$$p(x_j) = \frac{\sum_{i=1}^{n} r(x_i \mid x_j)}{n}, \text{ and} \tag{6}$$

$$r(x_i \mid x_j) = \begin{cases} 0 & \text{if } x_i \neq x_j \\ 1 & \text{if } x_i = x_j \end{cases}, \tag{7}$$

where i=1,2,..., i, j, ...n.

The kernel function, $K(x_i \mid x_j)$ is summed across all j unique observed scores weighted by their probability of occurrence, $p(x_j)$ and n is the sample size. The kernel function (Equation 5) used is the Gaussian normal distribution function, with a parameter, *h*, defining the "spread" of the kernel. Larger values of *h* result in a wider kernel, and thus produce a greater amount of smoothing. As *h* increases, the sampling error decreases, but the estimated distribution becomes more uniformly flat, resulting in increased bias.

Given the nature of the score distribution, it is likely that the distance between adjacent observed scores may differ significantly, depending on where the adjacent scores in the population density. For example, around the average, adjacent scores may be quite close, because the probability of a score being in that range is very high. However, extreme scores are much more rare, and the distance between adjacent extreme scores may be a tenth of a standard deviation or greater. For this reason, it

is appropriate to vary the bandwidth parameter, h, in the above kernel function to reduce "bumpiness" in the estimated density function. The optimal kernel bandwidth used for this analysis is the "Better Rule of Thumb" bandwidth defined by Härdle (1990, p. 91) as

$$h = \frac{4}{3} Min\left(\sigma, \frac{R}{1.34}\right) n^{-\frac{1}{5}}, \tag{8}$$

where s and R are the standard deviation and interquartile range of the observed scores, respectively. Cope and Kolen (1990) suggested that varying the kernel function for each observed data point could increase the accuracy of a kernel estimator. This adjustment allows the kernel to provide more smoothing where scores are sparse, while avoiding over-smoothing where the score distribution is denser. Variable kernels are described in Silverman (1986). The optimal kernel described above was varied according to the k^{th} root of the mean distance to the nearest neighbours for each observed score. The value of k was determined based on comparison of several values—in general, higher roots will result in a more uniform kernel, and lower roots will result in a more varied kernel. This adjustment produced the following function for h^*_{xj} for score x_j:

$$h^*_{xj} = h\left(\frac{x_{j+1} - x_{j-1}}{2}\right)^{\frac{1}{k}}. \tag{9}$$

In order to speed the algorithm, the density was defined discretely at 2000 sample points evenly distributed between arbitrary endpoints. Silverman (1986) recommended endpoints equal to the maximum and minimum observed scores, plus and minus $3h$. These values were used in the present study. Finally, a piecewise linear function was interpolated between the sample points and divided by its definite integral, limited by the endpoints defined above, to produce the final density estimate, $P(x_j)$.

Finite mixture models

The essential premise of a Gaussian mixture model is that any observed non-normal distribution can be described by a finite number of normal distributions. In many educational situations, this type of clustering is expected, given the influence of factors like institutional effectiveness and subpopulation membership on performance (Bryk & Raudenbush, 1992). In a most basic example, a distribution of test scores may have two distinct modes, corresponding to two systematically different subpopulations. If it were possible to determine which individuals belonged to each subpopulation, then it would be possible to estimate the parameters describing the normal distribution of each subpopulation separately. Adding the two components, weighted by their relative size, develops a parametric model of the overall population distribution. In more complex cases, where there are many distinct subpopulations, each normally distributed, it would be necessary to estimate the distribution of each subpopulation separately. Two types of mixture models were investigated in the present study: variable component models and fixed component models.

Catalogue no. 81-595-MIE2003005

Variable component mixture models

The full variable component model takes the form

$$P(x|\, W,M,S)_j = \sum_{i=1}^{k} w_i \left[\frac{1}{s_i \sqrt{2\pi}} e^{-\frac{1}{2}\left(s_i^{-1}(x-m_i)\right)^2} \right], \qquad (10)$$

where W is the vector of weights for the components, M is the vector of locations for the component, and S is the vector of standard deviations for each normally distributed component. The W, M and S vectors are all of length k, the number of components used in the mixture.

In practice, it is rare that the subpopulation membership is known for each individual. Furthermore, the number of subpopulations is also unknown. Thus, the estimation of a finite mixture model uses a maximum likelihood approach to estimate the parameters for each model with a specified number of components, as well as to evaluate estimated models with different numbers of components. The estimation of model parameter vectors (W,M,S) from a given number of components follows the Expectation-Maximization (EM) algorithm for finite mixtures (see, e.g., Dempster, Laird, and Rubin, 1977; Everitt and Hand, 1981).

Fixed Mixture Models

The family of fixed mixture models is identical to the variable-component method of Equation 17, with the exception that the number and location of the distributions is fixed to an arbitrary maximum. That is, the elements, m_i, of M for the fixed mixture models were defined by

$$m_i = (i+1)\frac{\text{Range}(X)}{k+1}. \qquad (11)$$

The remaining parameters (component variances and weights) are estimated by the EM algorithm used for the variable component models. This approach represents a compromise between the variable mixture model and the kernel density estimator. Although the parameter estimation process is mathematically identical, conceptually, the focus shifts from estimating the parameters of each individual component to producing a parametric maximum likelihood model of the full sample distribution. Thus, even though the locations of the components are fixed, any observed modality in the data can be modelled by changing the relative weightings and variances of adjacent components. A noted problem with estimating mixture models using iterative techniques is that there is always the possibility of components becoming singularities (Everitt and Hand, 1981). That is, as the location of a component converges onto the location of a single data point, its variance will go towards zero, since, theoretically, the most likely model to produce the data is a singular component for each observation. By fixing the locations vector, it may be possible to produce a parametric function that accurately describes the data density without concern for convergence on singularities.

4-Parameter Beta Distributions

The 4-parameter beta model is based on the 2-parameter Beta distribution, which has a lower bound of zero and an upper bound of one. The two parameters define the location, skewness and kurtosis, within the (0,1) domain. The 4-parameter Beta distribution adds an additional two parameters that are used to redefine the minimum and maximum limits of the score domain. The 4-parameter model is defined by

$$g(x|\alpha,\beta,l,u)=\frac{(-l+x)^{\alpha-1}(u+x)^{\beta-1}}{(x-l)^{\alpha+\beta-1}B(\alpha,\beta)}, \qquad (12)$$

where B(a,b), the beta function, is related to the gamma function, G(x) = (x-1)! (for x>0), by

$$B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \qquad (13)$$

The a and b parameters describe the location, skewness, and kurtosis of the distribution, given the domain. The other two parameters, l and u, describe the lower and upper limits of the domain, respectively. Estimation of the four parameters follows the method of moments algorithm for estimating three parameters of the 4-parameter Beta distribution, when the fourth parameter is specified (described in Johnson & Kotz, 1970, and amended by Hanson, 1991). The algorithm used in the current study requires specification of the lower limit parameter and estimates the remaining parameters accordingly, using the non-central moments of the sample data. The specified value of the lower limit was chosen to minimise the squared deviation of the higher moments of the fitted distribution from the higher moments of the sample.

3. Uncertainty in Transformed Scores: Measurement Error, Linking Error, and Sampling Error

Error in Scores For Individuals

Because inferences of this linking function are based on a sample, there is uncertainty regarding the true nature of the linking function, referred to here as *linking error*. Linking error describes how different the linked scores are expected to be from their true values, had students actually written both assessments. When describing the expected score equivalent for any individual, it is necessary to include this description of uncertainty. Moreover, the score for an individual on the original scale also contains uncertainty, described by the measurement error. Thus, when rescaling an individual score, two sources of error must be considered:

- measurement error
- linking error

Catalogue no. 81-595-MIE2003005

The total error variance for an individual score that has been rescaled is estimated by combining the linking error variance and measurement error variance:

$$\text{var}_{individual}[e_y x] = \text{var}_{linking}[e_y x] + \text{var}_{measurement}[e_y x]. \quad (14)$$

Linking Error

Linking errors for the results obtained in the present study are estimated using Lord's equation (see Kolen and Brennan 1995, p. 227) for estimating the standard error of equipercentile equating using the random group design. The final result of the score transformation is the product of many parameters, each estimated from the data with some sampling error. The final error component in the transformed score is the sum of the errors and their covariances introduced during the estimation of each parameter. Because the current study assumes a continuous, rather than discrete, score distribution, the elements in Lord's formula representing approximations of CDF and PDF values from discrete distributions have been replaced by their continuous distribution versions. The revised Equation (15) represents the uncertainty in estimating a score on scale Y from a given score x_t on scale X, $\text{var}_{linkng}[e_y x_t]$:

$$\text{var}_{linking}\left[\hat{e}_y x_t\right] = \left[PDF_y\left(\hat{e}_y x_t\right)\right]^{-2}\left\{\frac{CDF_x(x_t)\left[1-CDF_x(x_t)\right](n_x+n_y)}{n_x n_y} - \frac{\left[CDF_y\left(\hat{e}_y x_t\right)-CDF_x(x)\right]\left[CDF_x(x_t)-CDF_x(y)\right]}{n_y\left[PDF_y\left(\hat{e}_y x_t\right)\right]}\right\} \quad (15)$$

where $PDF_y(\hat{e}_y x_t)$ is the proportion of examinees estimated to have scale score $\hat{e}_y x_t$, and $CDF_x(x)$ and $CDF_y(\hat{e}_y x_t)$ are the proportions of examinees estimated to have scale scores at or below x and y, respectively. Kolen and Brennan (1995, p. 240), suggest that a tolerable linking error for interpretation of linked scores be less than 0.1 of a standard deviation. For both assessments considered in this study, the score standard deviation is 100.

Measurement Error

For the assessments that the methods explored in this study are expected to be applied, there are three main techniques of quantifying error of measurement:

- Each score is assigned a common measurement error, equal to the standard deviation of normally distributed error specific to the full test form;
- Each score is assigned a specific measurement error, equal to the standard deviation of a presumed normally distributed[5] error specific to the items each examinee responds to and his or her proficiency (e.g., FSA); and
- Each examinee is assigned several scores, which together represent the probability distribution of his or her proficiency (e.g., PISA).

Measurement error is expressed in terms of the scale of the original score. Consequently, in order to express measurement error of the transformed score, $e_y x_t$, the individual measurement variance errors of scale X must first be converted to scale Y by applying the linking function to the measurement error variance. For the first two scenarios, the variance error of measurement can be rescaled directly using

the following formula:

$$\text{var}_{measurement}[e_y x] = \text{var}(x)\left[\frac{(1-PDF_x(x))PDF_x(x)}{(1-PDF_y(e_y x))PDF_y(e_y x)}\right]^2. \quad (16)$$

Equation 16 describes the multiplication of the variance error of score x on scale X by the ratio of the distribution variance of score x to the distribution variance of the rescaled score, $e_y x$, on scale Y.

For the third scenario, each of the original imputations must be rescaled separately. The set of rescaled scores will represent the posterior density of proficiency for each examinee on the transformed scale. That is,

$$\text{var}_{measurement}[e_y(pv_{x1}, pv_{x2}, \ldots, pv_{xM})] \approx \left(1+\frac{1}{M}\right)\frac{1}{M-1}\sum_{i=1}^{M}\left(e_y pv_{xi} - \overline{e_y pv_x}\right)^2 \quad (17)$$

for the set of M plausible values (adapted from Mislevy, 1991), where $e_y pv_{xi}$ is used to denote the scale Y equivalent for ith plausible value drawn for an examinee on scale X, and $\overline{e_y pv_x}$ is the mean of the five scale Y equivalents. Alternately, the same results can be achieved by using the average of the plausible values as a single point estimate for each case. Estimating the measurement variance on the original scale (using Equation 2 with the original plausible values) and converting this variance with Equation 16 will produce approximately the same estimate of converted variance as Equation 17. However, any use of plausible values at the individual level will be inaccurate. Each plausible value represents a random draw from an individual posterior density, and their mean is an approximation to the mean of the posterior density. As this can be estimated directly, it is recommended that any rescaling for interpretation at the individual level should use directly estimated point estimates (e. g., maximum likelihood, weighted likelihood, *expected a posteriori*) and their corresponding parametric measurement errors.

Error in Estimated Sample Statistics

The purpose of this type of rescaling exercise, as previously described, is to produce group level statistics rather than individual estimates. In order to estimate group statistics from rescaled scores, three sources of error must be considered:

- measurement error
- linking error
- sampling error

The goal in this section is to define a method for combining the various sources of error to describing the uncertainty of a statistic estimated from the sample. The general form of the equation for estimating the total variance in a statistic is:

Variance = measurement variance + linking variance + sampling variance. (18)

Measurement and Linking Variance

The first two variance components can be estimated simultaneously using Equation 14, which combines measurement and linking variance at the individual level. The group combined variance is estimated by first recreating a proficiency density function for each individual based on an available point estimate and its combined errors of measurement and linking. These individual distributions will be summed for the entire group to produce a continuous function describing the distribution of sample scores:

$$P(e_y x) = \frac{1}{n}\sum_{i}^{n} p_i(e_y x)\,; \tag{19}$$

$$p_i(e_y x) = \left[\mathrm{var}_{individual}(e_y x_i)^{\frac{1}{2}}\sqrt{2\pi}\right]^{-1} e^{-\frac{1}{2}\left((e_y x - e_y x_i)/\mathrm{var}_{individual}(e_y x_i)^{\frac{1}{2}}\right)^2}. \tag{20}$$

Although the derivation of all commonly used statistics is outside the scope of this study, two examples of estimation will be presented: the sample mean and the proportion of the sample between two cut-points, along with their respective standard errors.

Combined Error for Means

The mean is defined as the first central moment of the continuous function, $P(e_y x)$, which is defined as the sum of each observed score value, weighted by its probability of occurence. The sample distribution is equal to the sum of the individual distributions weighted by their individual probability. Individual errors are assumed independent; hence, the covariances between scores are assumed to equal 0. Furthermore, each individual score has a distinct combined measurement and linking error. Therefore, the combined measurement and linking variance of the mean is equal to the sum of the variances, weighted by the square of their individual probability:

$$\mathrm{var}_{meas.link.}\left(\overline{e_y x}\right) = \sum_{i}^{n} \mathrm{var}_{individual}(e_y x)\, p^2\,. \tag{21}$$

Combined Error for Proportions

The calculations for error of estimate of proportions are similar. However, each individual contribution to the sample statistic is calculated as the definite integral, $p_i(e_y x \mid L1, L2)$, which describes the probability that the ith individual's true score lies between the limits of integration, L1 and L2:

$$\pi_i(e_y x \mid L1, L2) = \int_{L1}^{L2} p_i(e_y x)\,\partial e_y x\,. \tag{22}$$

The sample statistic for the proportion of individuals with scores between these cut points, $\Pi(e_y x|\, L1, L2)$, is estimated as the sum of the n definite integrals, weighted by their probabilities:

$$\Pi(e_y x|L1, L2) = \sum_i^n \pi_i (e_y x|L1, L2)\, p \ . \qquad (23)$$

Under the assumption of independence of observations, the combined measurement and linking error of the proportion defined in Equation 23 is estimated as the sum of variances of each of the n definite integrals, weighted by the square of their probability:

$$\text{var}[\Pi(e_y x|l1, l2)] = \sum_i^n \pi_i (e_y x|l1, l2)[1 - \pi_i (e_y x|l1, l2)]\, p^2 \ . \qquad (24)$$

Sampling Variance

The sampling variance of the statistic is estimated independently of measurement and linking error. However, estimation of sampling variance typically depends on some knowledge and consideration of the sample design. For the FSA, reporting of provincial results assumes that the students represent a census, and no sampling error is calculated. For simple random samples, it is expected to be proportional to the sample standard deviation.

For complex samples, typically an appropriate estimation strategy based on Taylor Series approximation or replication is used. For example, sampling variance estimation for PISA analysis uses a balanced repeated replicate (BRR) method (see the *PISA 2000 Technical Report,* 2002). The error variance estimates from each of the above procedures are summed to produce the overall parameter variance. Thus, Equation 18 is operationalized in this situation as:

$$\text{var}_{total}(\overline{e_y x}) = \text{var}_{meas.link.}(\overline{e_y x}) + \text{var}_{sampling}(\overline{e_y x}) \ . \qquad (25)$$

References

Braun, H.I. & Holland, P.W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland and D. B. Rubin (Eds.), *Test Equating* (pp. 9-49). New York: Academic.

British Columbia Ministry of Education (2000, 2001, 2002). Interpreting and communicating British Columbia Foundation Skills Assessment Results. Available: http://www.bced.gov.bc.ca/assessment/fsa/

Bryk, A. S. , & Raudenbush, S. W. (1992). *Hierarchical Linear Models.* California:Sage.

Cartwright, F. (in press). Equipercentile methods of linking regional and inter-regional assessments.

Cope, R. T. , & Kolen, M. J. (1990). *A Study of Methods for Estimating Distributions of Test Scores.* Iowa: American College Testing.

Dempster, A., Laird, N., & Rubin, D. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39 (Series B),* 1-38.

Everitt, B. S. , & Hand, D. J. (1981). *Finite Mixture Distributions.* New York: Chapman and Hall Ltd.

Hanson, B. (1990). *An Investigation of Methods for Improving Estimation of Test Score Distributions.* Iowa: American College Testing.

Hanson, B. (1991). *Method of Moments Estimates for Four-parameter Beta Compound Binomial Model and the Calculation of Classification Consistency Indexes.* Iowa: American College Testing.

Härdle, W. (1990). *Smoothing Techniques with Implementation in S.* New York: Springer-Verlag.

Johnson, J. L., & Kotz, S. (1970). *Continuous Univariate Distributions 2.* Boston: Houghton Mifflin Company.

Kolen, M. J., & Brennan, R. L. (1995). *Test Equating: Methods and Practices.* New York: Springer.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6,* 83–102.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel forms reliability. *Psychometrika, 48,* 233-245.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56,* 177-196.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: Policy Information Center, Educational Testing Service.

OECD (2001). *Knowledge and Skills for Life. First Results from PISA 2000.* Organisation for Economic Cooperation and Development. Available: http://www.pisa.oecd.org/

PISA Canada (2001). *Measuring Up: The Performance of Canada's Youth in Reading, Mathematics and Science. OECD PISA Study – First Results for Canadians aged 15.* Human Resources Development Canada, Council of Ministers of Education, Canada, and Statistics Canada. Available: http://www.cmec.ca/pisa/2000/indexe.stm

PISA 2000 Technical Report. (2002). R. Adams and M. Wu, Ed. Paris: Organisation for Economic Co-operation and Development.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* New York: Chapman and Hall Ltd.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54,* 427-450.

Endnotes

1 The G-7 consists of seven industrialized countries: Canada, France, Germany, Italy, Japan, United Kingdom, and the United States.

2 Considerations and limitations for interpreting the results of linkages are explained in detail in Mislevy (1992) and Linn (1993).

3 More information on FSA can be found at www.bced.gov.bc.ca/assessment/fsa/.

4 More information on PISA can be found at www.pisa.oecd.org.

5 Error is not always normally distributed, because fixed test forms are likely to provide more information around average ability levels (see Lord, 1983; Warm, 1989).

Culture, Tourism and the Centre for Education Statistics

Research Papers

Cumulative Index

Statistics Canada's **Division of Culture, Tourism and the Centre for Education Statistics** develops surveys, provides statistics and conducts research and analysis relevant to current issues in its three areas of responsibility.

The **Culture Statistics Program** creates and disseminates timely and comprehensive information on the culture sector in Canada. The program manages a dozen regular census surveys and databanks to produce data that support policy decision and program management requirements. Issues include the economic impact of culture, the consumption of culture goods and services, government, personal and corporate spending on culture, the culture labour market, and international trade of culture goods and services. Its analytical output appears in the flagship publication *Focus on Culture* (www.statcan.ca/english/IPS/Data/87-004-XIE.htm) and in *Arts, culture and recreation – Research papers*.

The **Tourism Statistics Program** provides information on domestic and international tourism. The program covers the Canadian Travel Survey and the International Travel Survey. Together, these surveys shed light on the volume and characteristics of trips and travellers to, from and within Canada. Its analytical output appears in the flagship publication *Travel-log* (www.statcan.ca/english/IPS/Data/87-003-XIE.htm) and in *Travel and tourism – Research papers*.

The **Centre for Education Statistics** develops and delivers a comprehensive program of pan-Canadian education statistics and analysis in order to support policy decisions and program management, and to ensure that accurate and relevant information concerning education is available to the Canadian public and to other educational stakeholders. The Centre conducts fifteen institutional and over ten household education surveys. Its analytical output appears in the flagship publication *Education quarterly review* (www.statcan.ca/english/IPS/Data/81-003-XIE.htm), in various monographs and in *Education, skills and learning – Research papers* (www.statcan.ca/english/IPS/Data/81-595-MIE.htm).

Following is a cumulative index of Culture, Tourism and Education research papers published to date

Arts, culture and recreation – Research papers

Forthcoming

Travel and tourism – Research papers

Forthcoming

Education, skills and learning – Research papers

81-595-MIE2002001	Understanding the rural-urban reading gap
81-595-MIE2003002	Canadian education and training services abroad: the role of contracts funded by international financial institution.
81-595-MIE2003003	Finding their way: a profile of young Canadian graduates
81-595-MIE2003004	Learning, Earning and Leaving – The relationship between working while in high school and dropping out
81-595-MIE2003005	Linking provincial student assessments with national and international assessments